植 物 分 类 学 报 31 (6): 578—586 (1993)
Acta Phytotaxonomica Sinica

分支分类学的一种计算方法——最小平行进化法

徐克学

(中国科学院植物研究所系统与进化植物学开放研究实验室，北京 100093)

AN ALGORITHM FOR CLADISTICS——METHOD OF MINIMAL PARALLEL EVOLUTION

XU KE-XUE

(Laboratory of Systematic and Evolutionary Botany, & Herbarium, Institute of Botany, Chinese Academy of Sciences, Beijing 100093)

Abstract The paper presented here is concerned with the numerical cladistics. In consideration of the fact that the parallel evolution has close relation to the length of evolution graph, a new method of reconstructing evolutionary tree has been developed for the application and practice of cladistics.

The procedure of the algorithm of the new method presented in Table 1 is similar to the method described in paper "An algorithm for cladistics ——method of maximal same step length".

An essential step of the algorithm is how to decide the coefficient between two cladistic units (CTUs). A coefficient called parallel evolutionary coefficient between CTUp and CTUq is defined as follows:

$$S_{pq}=\sum_{j=1}^{n}\sum_{\substack{i=1\\ i\neq p,\, q}}^{t} E(p,\ q,\ i,\ j),$$

where the j is code of CTU and the i is code of character; $E(p,\ q,\ i,\ j)$ is a function given by following expression:

$$E(p,q,i,j)=\begin{cases}\min\ (X_{ij},X_{pj})+\min\ (X_{ij},X_{qj})-2\min\ (X_{pj},X_{qj}) & \text{as } X_{ij}>\min\,(X_{pj},X_{qj})\\ 0 & \text{otherwise.}\end{cases}$$

where the X_{ij} is the ith row (CTU) jth colunm (Character) element of the data matrix.

Because the method of minimal parallel evolution is closely related to the length of evolutionary graph, it is superior to the method of maximal same step

本稿于1989年8月31日收到、经审稿，笔者实践，至1992年3月编委会再审稿后决定采用。

length. A simple datum as an example for comparison shows that the method of minimal parallel evolution can arrive at a better result.

But in some cases, we may combine one method with another and thus the coefficient should take following form:

$$S\ (S)_{ij} = M \cdot S\ (C)_{ij} - N \cdot S\ (P)_{ij}\ ,$$

in which $S\ (C)_{ij}$ and $S\ (P)_{ij}$ are the same step coefficients and the parallel evolution coefficient respectively, and the M and N are positive integers as a weight number being given in advance.

Key words Cladistics, Method of minimal parallel evolution

摘要 本文是"分支分类的一种计算方法——最大同步法"一文的姐妹篇。两种方法运算过程基本相同，不同之处乃是最小平行进化法利用平行进化的概念，首先确立两个分支单位相结合时产生平行进化的步数，即平行进化系数的计算公式，对所有待结合分支单位间计算平行进化系数。然后根据俭约性原理，要获得最俭约演化树谱图，应该尽可能减少平行进化，也就是说在选择结合的分支单位时，选择平行进化系数最小者优先结合。于是建立起一种新的分支分类运算方法。两种方法的思路完全不同，从原理上讲对某些数据，最小平行进化法优于最大同步法，但后者运算量较大。如果将两种思路兼顾，可以得出由这两种方法相结合而产生平行同步综合法。桔梗科6个种的数据作为例子进行运算说明。

关键词 分支分类学；最小平行进化法

引　　言

本文作者于1989年发表了"分支分类的一种计算方法——最大同步法"一文，受到国内外学者的关注和引用，实践和应用说明这是一个成功的分支分类运算方法。方法之所以成功，首先在于有扎实的理论基础，它基于"演化集合论"。

其次，最大同步法属于合理方法[1)]。所谓合理方法，是指对完全和谐的原始数据能获得分支分类问题的最优解。

再次，该方法把分支谱系原理与聚合的表征分类运算格式结合在一起。大大减化了计算机编写程序工作，为生物学家的应用提供了方便。

鉴于以上情况，笔者认为有必要在最大同步法的基础之上，继续进行研究，开拓出更多的分支分类新方法。本文作为最大同步法一文的姐妹篇奉献于读者。

本文提出的最小平行进化法与最大同步法相似，都以"演化集合论"为基础，运算格式基本相同。现在仅就二者的区别，在此进行介绍和讨论。

在许多实际问题中，最大同步法虽然能获得比较满意的谱系图，但是它仅仅考虑到把相同性状的进化状态都表现在同一分支线上。没有考虑到获得最短演化路径的另一个方面，即平行进化的关系。下面一组简单数据和它的计算结果显示了最大同步法的不尽完善之处：

1）徐克学：数量分类学．科学出版社（待发表）．

		性状 characters					
		1	2	3	4	5	6
	1	1	2	0	2	0	0
OTUs	2	1	2	2	0	0	0
	3	0	0	0	2	0	1
	4	0	0	2	0	1	0

这个数据的最大同步法计算结果在图1中，演化长度为13。演化图中性状3和性状4共产生了4步平行进化，导致演化长度的浪费。实际上，采用新的方法计算，演化长度还可以减少。

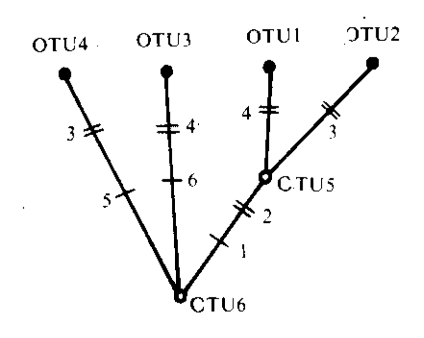


图1 最大同步法的一个计算结果

Fig. 1 A computation result of the method of maximal same step length

思 路

此方法是就前文的数据改换思路并进行深入研究的结果（为了节省篇幅，本文引用的名词、概念和符号，凡前文已经详细介绍过不再重复，必要时请参阅该文)。按照最大同步法，将每一对OTUs的同步系数计算出来，排成如下形式:

	1				
	2	3			
OTU_i	3	2	0		
	4	0	2	0	
		1	2	3	4
		OTU_j			

同步系数
same step coefficients
$S(C)_{ij}$

其中同步系数$S_{21}=3$最大。最大同步法的基本思想就是为了把相同性状的进化尽可能都表现在同一条分支线上。为了表述方便，在演化图中以分支点X_i代表分支单位CTU_i或分类单位OTU_i。首先将分支单位CTU_2和CTU_1相结合，从而在演化图(图1)中获得从分支点X_5分别到分支点X_1和X_2的分支线。然而这样的结合是不是合理呢？检查整个演化图就会发现分支线(X_5，X_1)与分支线(X_6，X_3)在性状4具有2步平行进化，性

状4在不同方向上重演相同的进化导致演化步长使用的浪费。同样情形也发生在分支线(X_5, X_2)与(X_6, X_4)对性状3具有2步平行进化。因而分支单位CTU_1与CTU_2的结合共产生了4步平行进化。演化图长度13步比最短演化步长$L_{min}=9$步高出4步，恰恰就是这种不合理的结合造成。由此给我们以启示，在构造演化图时，同步系数不是分支点相结合的最主要依据，为了获得较短的演化图长度，亦可以从平行进化的角度进行考虑。

将两分支单位结合后，在两个分支线上能够产生平行进化的总步长愈长，演化图长度亦愈长。像前面CTU_1与CTU_2的结合就产生了4步平行进化。很显然，在构造演化图时，应该使每次相结合的分支单位产生平行进化的长度愈小愈佳。这就是我们构造演化图的新依据。因此我们十分关心对产生平行进化的步长进行估计，它是我们引导出一种新的分支分类方法的关键，为了把新方法阐述清楚，现从最原始步骤做起。

如果被研究的对象是t个OTUs，观察记录的n个性状，全部性状状态的编码数排成t行n列原始数值矩阵，

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{t1} & x_{t2} & \cdots & x_{tn} \end{bmatrix}$$

矩阵中元素x_{ij}表示第i分类单位(行)在第j性状(列)的性状状态编码值。性状进行编码时，使所有的性状编码值都取非负整数，从0开始，依演化的次序而增加。编码值为0的状态是最原始的状态。为了表述方便，可采用向量或演化图中的分支点表示分类单位或分支单位，如第i分类单位可表示成$X_i=(x_{i1}\ x_{i2}\cdots x_{in})$。

分支分类的目的就是要从已经获得的原始数值矩阵，去推测符合真实情况的演化图。考虑生物系统发育的规律性，我们对构造的演化图作如下假设:

1. 演化图在任何演化路径上没有回路;
2. 演化图中任意两分化的分支单位，不可能再融合而产生网状进化;
3. 整个演化图是单源的，即都由一个共同的祖先演化而来;
4. 演化图的演化方向与性状的进化保持一致;
5. 演化图遵从最短演化路径原理，即代表生物演化关系的演化图，在一切可能构造的图中演化长度取最小值。

有了前面一系列的准备工作，现在开始着手介绍如何构造演化图。我们仍然采取与最大同步法相类似的运算过程，不断地将分支单位相结合，直到所有的分类单位都被结合在一棵树状图中。在构造演化图的过程中，从众多的分支单位中，选择哪一对先结合? 这时，我们不再考虑同步系数，而是从如何得到最小平行进化来考虑。

如果将分支单位$X_p=(x_{p1}\ x_{p2}\cdots x_{pn})$ 与$X_q=(x_{q1}\ x_{q2}\cdots x_{qn})$ 相结合，为了寻找一个与平行进化长度有关的指标，先定义

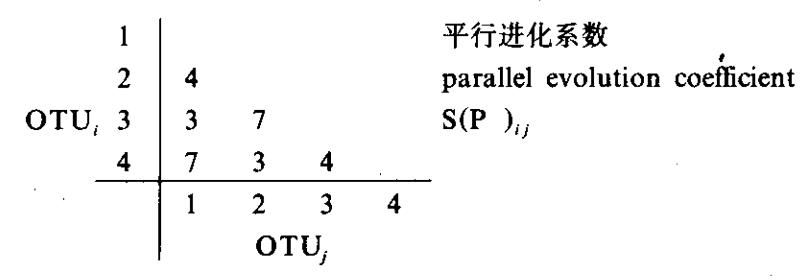
$$E(p,q,i,j)=\begin{cases}\min(x_{ij}, x_{pj})+\min(x_{ij}, x_{qj})-2\min(x_{pj}, x_{qj}) & \text{当 } x_{ij}>\min(x_{pj}, x_{qj}) \\ 0 & \text{其它情形}\end{cases} \tag{1}$$

式中 i ($i \neq p$, q)和 j ($j=1, 2, \cdots, n$) 分别表示分支单位和性状的标号。若两分支单位 CTU_p 与CTU_q 相结合，为了估计CTU_p和 CTU_q 所在两分支线上可能产生平行进化的长度，先固定j，对于 i 在演化图中从最原始的祖先到分类单位 OTU_i 有一条演化路径，值 $E(p, q, i, j)$ 就是该演化路径上，与 CTU_p 和 CTU_q 所在的两分支线形成的平行进化长度。让标号i 跑遍所有的分支单位(除去 p 和 q)再让 j 跑遍所有的性状，对 $E(p, q, i, j)$ 求和

$$S_{pq}=\sum_{j=1}^{n}\sum_{\substack{i=1\\ i\neq p,q}}^{t} E(p, q, i, j). \tag{2}$$

这就是分支单位 CTU_p 与 CTU_q 相结合，由此而可能产生平行进化的一个估计值，称之为分支单位p 与 q 的平行进化系数。平行进化系数是我们构造演化图的依据。在构造演化图的过程中，应该取平行进化系数最小的分支单位结合在一起。

以前述数据为例，计算每一对分类单位的平行进化系数，排成矩阵形式如下：



	1				
	2	4			
OTU_i	3	3	7		
	4	7	3	4	
		1	2	3	4
			OTU_j		

平行进化系数
parallel evolution coefficient
$S(P)_{ij}$

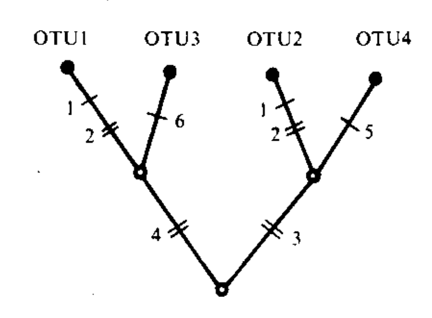


图2 最小平行进化法的一个计算结果

Fig. 2 A computation result of the method of minimal parallel evolution

矩阵中 $S(P)_{31}=S(P)_{42}=3$ 最小，应该首先将分类单位 3 和 1 或者分类单位4 和 2 相结合，而不应该将分类单位1 与 2 相结合。

有了平行进化系数，即可着手构造演化图。

运 算 步 骤

构造演化图的运算方法与最大同步法十分类似，它亦由多次循环过程完成。每次循环的具体运算步骤列出如下：

1. 按性状状态的进行次序进行性状编码，每个性状的最原始状态取 0 值，其它状

态依进化次序从小到大取非负整数，得 t 行(OTU)n 列(性状)原始数值矩阵，置数据矩阵中。

2. 利用公式(2)和(1)计算数据矩阵中所有分支单位间的平行进化系数 S_{ij} $(i \neq j)$，置系数矩阵中。

3. 从系数矩阵中找出平行进化系数最小值。假如就是 S_{pq}，由此确定把分支单位 CTU_p 与 CTU_q 相结合。若有两个以上平行进化系数达到最小值，可任择一个执行。

4. 求出分支点 X_p 与 X_q 的最近共同祖先 X_r 的性状分量值，该值是分支点 X_p 与 X_q 相应分量的最小值：

$$x_{ri} = \min\ (x_{pi},\ x_{qi}) \quad (i=1,\ 2,\ \cdots,\ n) \tag{3}$$

然后从数据矩阵中删除分支单位 CTU_p 和 CTU_q 的数据，补充以新的分支单位 CTU_r。矩阵的分支单位(行)数比原来减少1。

5. 在演化图上作出从分支点 X_r 到分支点 X_p 和 X_q 的分支关系，记下该分支线的演化长度和产生演化的性状。

若数据矩阵的分支单位(行)数 ≥2，则转向步骤2进入下一次循环运算。否则运算结束。

最后检查演化图中是否出现演化长度为0的"分支线"。若有，将完全相同的起点与终点重合，取消演化长度为0的"分支线"。

以桔梗科6个种的数据为例，演算过程列于表1。运算结果画出演化图(见图3)。

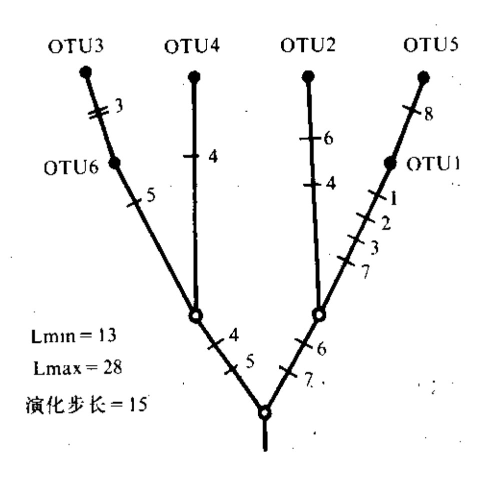


图3　桔梗科6个种数据分支树谱图

Fig. 3　Cladogram of data of 6 species from Campanulaceae

OTU1　*Codonopsis pilosula* (Franch.) Nannf. 党参

OTU2　*Platycodon grandiflorus* (Jacq.) A. DC. 桔梗

OTU3　*Adenophora pereskiifolia* (Fisch.) G. Don 轮叶沙参

OTU4　*Adenophora remotiflora* Miq. 荠苨

OTU5　*Codonopsis lanceolata* Benth. et Hook. f. 羊乳

OTU6　*Adenophora polyantha* Nakai 石沙参

表1 桔梗科6个种最小平行进化法分支分类运算过程

Table 1 A computing process of cladistic taxonomy by method of minimal parallel evolution on the data of 6 species from Campanulaceae

循环次数 Times of cycle	分支单位编码 No. of CTU	系数矩阵 Coefficient matrix	性 状 Characters:	数据矩阵 Data matrix 1	2	3	4	5	6	7	8
I	1	X		1	1	1	0	0	1	2	0
	2	8 X		0	0	0	1	0	2	1	0
	3	13 9 X		0	0	2	1	2	0	0	0
	4	14 6 3 X		0	0	0	2	1	0	0	0
	5	0 8 13 14 X		1	1	1	0	0	1	2	1
	6	15 7 2 1 15 X		0	0	0	1	2	0	0	0
II	7	X		1	1	1	0	0	1	2	0
	2	4 X		0	0	0	1	0	2	1	0
	3	8 6 X		0	0	2	1	2	0	0	0
	4	8 4 2 X		0	0	0	2	1	0	0	0
	6	9 5 11 X		0	0	0	1	2	0	0	0
III	7	X		1	1	1	0	0	1	2	0
	2	2 X		0	0	0	1	0	2	1	0
	8	5 3 X		0	0	0	1	2	0	0	0
	4	5 3 0 X		0	0	0	2	1	0	0	0
IV	7	X		1	1	1	0	0	1	2	0
	2	1 X		0	0	0	1	0	2	1	0
	9	3 2 X		0	0	0	1	1	0	0	0
V	10			0	0	0	0	0	1	1	0
	9			0	0	0	1	1	0	0	0
VI	11			0	0	0	0	0	0	0	0

说 明

第I次循环运算:

先根据公式(1)和(2)从数据矩阵计算平行进化系数。例如计算 S_{12}:

当 $j=1$ 时，$E(1, 2, 5, 1)=1$，其它 i 值 $E(1, 2, i, 1)$ 均为0;

当 $j=2$ 时，同前。

当 $j=3$ 时，$E(1, 2, 3, 3)=E(1, 2, 5, 3)=1$，其它 i 值 $E(1, 2, i, 3)$ 均为0。

……

当 $j=8$ 时，$E(1, 2, i, 8)$ 均为0。

因而有 $S_{12}=1+1+2+3+0+0+1+0=8$。

然后根据最小平行进化系数 $S_{15}=0$，确定 CTU_1 和 CTU_5 相结合，二者的最近共同祖先是 CTU_7。利用公式(3)求出 X_7 的性状分量值，

$$X_7=(1\ 1\ 1\ 0\ 0\ 1\ 2\ 0)。$$

在演化图上作分支点 X_1，X_5 和 X_7 以及相应的分支线，表示从 CTU_7 进化到 CTU_1 和 CTU_5。

第 II 次循环运算:

以 X_7 取代 X_5 和 X_1，置 X_7 的性状分量值于数据矩阵中，得新的数据矩阵，并计算所有分支单位间的平行进化系数。

依照本次循环最小平行进化系数 $S_{63}=S_{64}=1$，不妨取 CTU_6 与 CTU_3 相结合，最近共同祖先为 CTU_8，找出性状分量值 $X_8=(0\ \ 0\ 0\ 1\ 2\ 0\ 0\ 0)$。

在演化图中补充相应的分支点和分支线，表示从 CTU_8 分别演化到 CTU_6 和 CTU_3

第 III 次循环运算:

以 X_8 取代 X_6 和 X_3，置 X_8 的性状分量值于数据矩阵中，得新数据矩阵。计算平行进化系数，得系数矩阵。最小平行进化系数 $S_{48}=0$，确定 CTU_4 和 CTU_8 相结合，最近共同祖先为 CTU_9。求出性状分量值 $X_9=(0\ 0\ 0\ 1\ 1\ 0\ 0\ 0)$。

在演化图上作出从分支点 X_9 到 X_4 和 X_8 的演化关系。

第 IV 次循环运算:

以 X_9 取代 X_4 和 X_8，置 X_9 的性状分量值于数据矩阵中，得新数据矩阵。计算 CTU_9 与 CTU_7，CTU_2 的平行进化系数，得系数矩阵。最小平行进化系数是 $S_{72}=1$，确定 CTU_7 与 CTU_2 相结合，最近共同祖先为 CTU_{10}。求出性状分量值 $X_{10}=(0\ 0\ 0\ 0\ 0\ 1\ 1\ 0)$。

在演化图上作出从分支点 X_{10} 到 X_7 和 X_2 的演化关系。

第 V 次循环运算:

以 X_{10} 取代 X_7 和 X_2，置 X_{10} 的性状分量值于数据矩阵中，新数据矩阵仅剩下两个分支单位，无选择余地，必然将两分支单位 CTU_9 与 CTU_{10} 相结合成 CTU_{11}，求出性状分量值 $X_{11}=(0\ 0\ 0\ 0\ 0\ 0\ 0\ 0)$。

在演化图上作出从分支点 X_{11} 到 X_9 和 X_{10} 的演化关系。

第 VI 次循环运算:

以 X_{11} 取代 X_9 和 X_{10}，数据矩阵剩下待处理的分支单位个数已小于 2，运算结束。

在演化图中将长度为 0 的分支线(X_8，X_6)和(X_7，X_1)取消，$X_8=X_6$，$X_7=X_1$，得演化图(图 3)。

讨　　论

许多实际的例子已说明最大同步法和最小平行进化法都能获得较好的分支演化图。至于这两种方法的比较，本文所举桔梗科 6 个种的例子，这两种方法的运算结果完全相同，此例尚不能显示何者较优。在此，先从方法的基本原理进行分析。

我们构造的演化图如果没有性状演化的逆转，通常演化长度与平行进化长度有下面的关系。

演化图长度 = 平行进化长度 + L_{min}。在此 $L_{min}=\sum_{i=1}^{n}\max_{k=1,2,\cdots,t}(x_{ki})$，对于给定的一

组原始数据，L_{min} 是常数。这个关系表示演化图中平行进化的步长与演化图的长度具有直接联系。因而最小平行进化法比最大同步法能更直接地体现最短演化路径原则。

从本文最初所举的数据能更清楚地说明这一点。前面已就这个数据计算了全部分类单位之间的同步系数和平行进化系数。令 $S(C)_{ij}$ 表示同步系数，$S(P)_{ij}$ 表示平行进化系数。在同步系数矩阵中 $S(C)_{12}=3$ 取到最大值，按照最大同步法，首先应将 OTU_1 与 OTU_2 相结合，但是平行进化系数值 $S(P)_{12}=4$，预示这种结合可能导致更多的平行进化。再察看平行进化系数矩阵中，$S(P)_{31}=S(P)_{42}=3$ 取到最小值，按照最小平行进化法应先将 OTU_3 与 OTU_1 或者 OTU_4 与 OTU_2 相结合，计算的结果(图 2)演化图长度比最大同步法优出 1 步。

同步系数虽然不及平行进化系数那样与演化图有较直接的联系，但是它毕竟是从事物另一个方面引导构造演化图的方法。在实际运算中、为了更全面，更周密地考虑问题，可以将两种方法结合使用。这种结合有两种方式。

第一种方式，以最小平行进化法为主，只是在构造演化图的过程中，如果出现多个最小平行进化系数，可选择同步系数最大者结合。

例如在表 1 所列的运算过程中，第 II 次循环出现了两个最小平行进化系数$S(P)_{36}=S(P)_{46}=1$。二者之中，取哪一对分支单位相互结合呢？ 计算同步系数$S(C)_{36}=3$，$S(C)_{46}=2$，进行比较有 $S(C)_{36}>S(C)_{46}$。应取 CTU_3 与 CTU_6 相结合。

第二种方式　设计新的系数，称为综合系数

$$S(S)_{ij}=MS(C)_{ij}-NS(P)_{ij} \tag{4}$$

式中$S(C)_{ij}$ 和 $S(P)_{ij}$ 分别是第 i, j 分支单位的平行进化系数和同步系数；M，N 作为权数事先给定，M 和 N 取正整数。不同的权数得到不同的综合系数。采用综合系数之后，构造演化图的步骤除第 3 步的最小值改为最大值外，其余都与最小平行进化法相同。

这两种方式引出的分支分类计算方法，可以称为平行同步综合法。在研究工作中选取适当的平行同步综合法，可望比单纯使用最小平行进化法或者最大同步法获得更加满意的效果。

本文所提出的方法，最小平行进化法连同各种形式的平行同步综合法，与最大同步法一样保持了易于计算机程序化的优点。请参看关于最大同步法一文的讨论。在设计计算机程序时，只要在系数计算等个别地方稍加更换，就可以从一种方法改变成另一种方法，我们可以将分支分类的最小平行进化法、最大同步法和各种形式的结合法，连同表征分类的多种分类方法都完成于同一个计算机程序之中。这将给分类系统学的研究带来更多的方便。

参 考 文 献

[1] 徐克学，1989：分支分类的一种计算方法——最大同步法。植物分类学报，27(3):232—239。
[2] 徐克学，1992：生物演化的数学模型。生物数学学报 7 (3)：92—97。